UPC Speech Activity Detector in RT06 Evaluation

Dušan Macho, Andrey Temko, Climent Nadeu

TALP Research Center, UPC Barcelona Spain

Rich Transcription 2006 Spring Meeting Recognition Evaluation Workshop

Bethesda, May 3-4, 2006

Presenting

- ❑ UPC Speech Activity Detector (SAD) for interactive meeting/lecture smart-room scenario in RT06 evaluation
- ❑ CHIL project – strong emphasis on unobtrusive and online technologies and demos – low-delay and real-time
- ❑ Aimed for several technologies (SID, SLOC, AED, ASR) and services
- ❑ Unobtrusive far-field microphone setup is assumed – SDM, MDM

- ❑ Three new SAD features added with respect to our previous work
- ❑ Two alternative classifiers have been tested in addition to Decision Tree

Features

ldam

30ms/10ms frame length/shift

Frequency Filtering (FF): filter $h(k)$={1, 0, -1} => static FF

Time derivatives: ΔFF, ΔΔFF, ΔlogE appended to static FF (16+16+16+1=49)

LDA: 49-element FF vector is reduced to 1-element scalar ldam

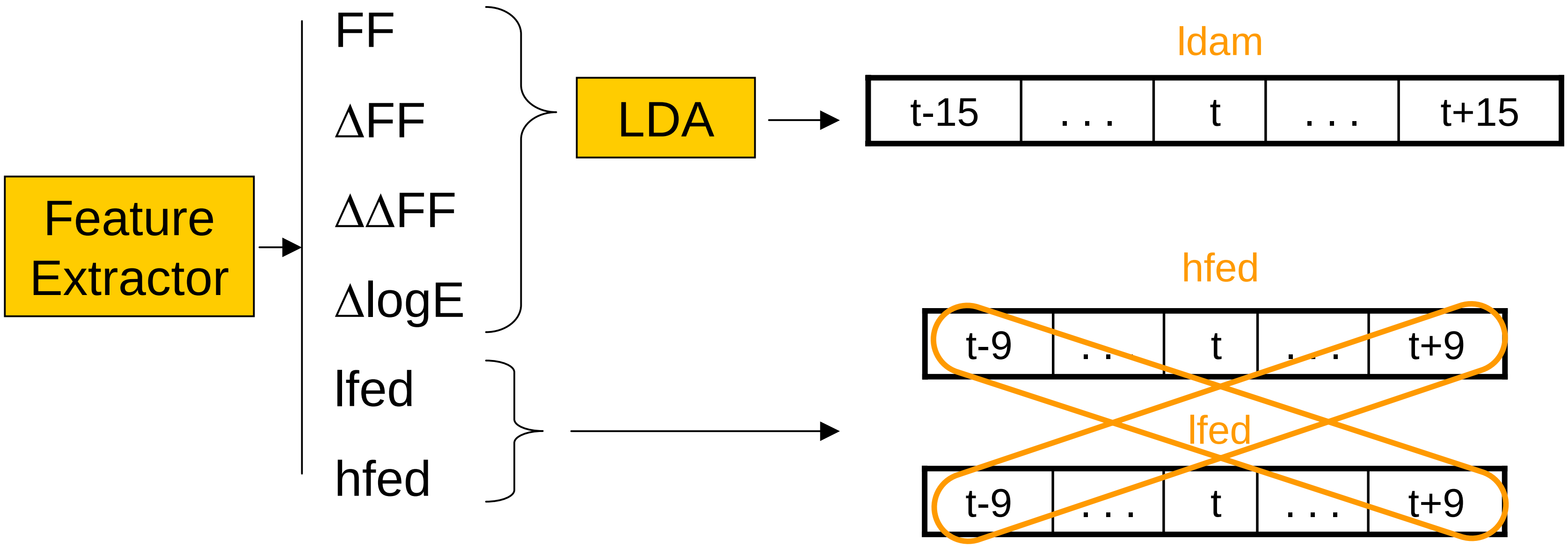
lfed, hfed

$$E_l(t) = \log\left(\sum_k S(k,t)\right)$$

where k correspond to 0.4-1.2kHz and 4.5-6.5 kHz for lfed and hfed, respectively

$$dE_l(t) = \frac{1}{60}\sum_{i=-4}^{4} i \cdot E_l(t+i) \qquad lfed(t) = \frac{1}{5}\sum_{i=-2}^{2} abs(dE_l(t+i))$$

Features, cont.



Feature Extractor

FF

ΔFF

ΔΔFF

ΔlogE

lfed

hfed

LDA

ldam

t-15 . . . t . . . t+15

hfed

t-9 . . . t . . . t+9

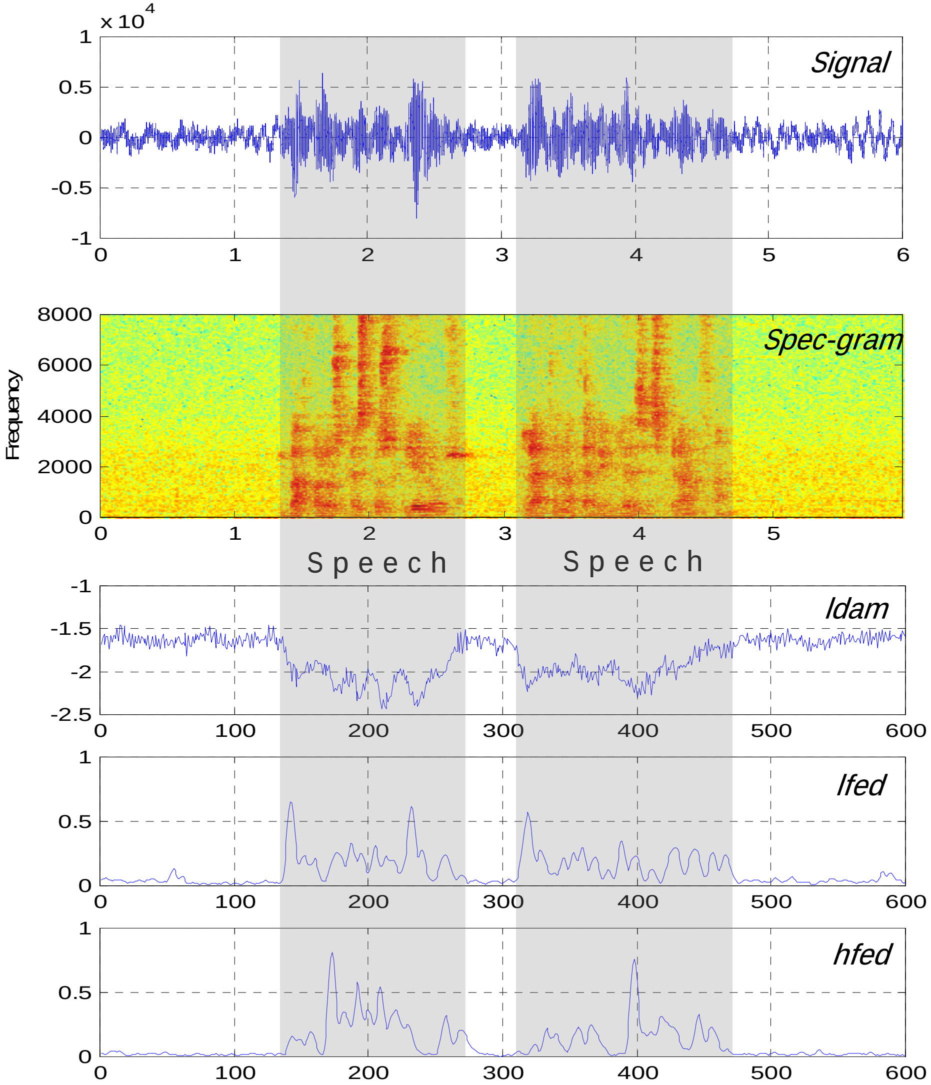
lfed

t-9 . . . t . . . t+9

$$\text{xfed}(t) = 1/2 * \left([\text{hfed}(t-9)*\text{lfed}(t+9)]^{1/2} + [\text{hfed}(t+9)*\text{lfed}(t-9)]^{1/2} \right)$$

Feature vector = ldam(t-15, t-10, t-6, t-3, t, t+3, t+6, t+10), lfed(t), hfed(t), xfed(t)

Features, cont.



Classifier

Gaussian Mixture Model (GMM)

❑ 32 mixtures for both Speech and Non-Speech with diagonal covariance matrix

❑ 20 iterations of EM algorithm for Gaussian mixture model training

❑ Classifier used in systems submitted for both “confmtg” and “lectmtg” tasks

Support Vector Machine (SVM, Andrey Temko)

❑ Training data set reduced to 19 thousand by using fast Proximal SVMs

❑ Gaussian kernel; parameters set via 5-fold cross-validation on the reduced training data

RT06 Post-processing

sdm

- 11 frame majority voting along time
- Addition of 0.2s at the beginning and the end of each speech segment

mdm

- sdm SAD for each channel (without post-proc.)
- Majority voting for each frame using info from several channels
- 11 frame majority voting along time
- Addition of 0.2s at the beginning and the end of each speech segment

Training data

Training needed for LDA and classifier

- confmtg: SPEECON and RT05 meeting
- lectmtg: above and CHIL

Database	SPEECON	RT05 meetings	CHIL
Language	Spanish	English	English
Type	Single utterances	Meeting	Lecture
Microphone	2-3m in front of speaker	On the table	On the table
Signal	16 kHz, 16 bit	16 kHz, 16 bit	16 kHz, 16 bit

Performance

Metrics
(RT06)

NIST = Duration of Incorrect Decisions / Duration of All Speech

Missed Spkr = Missed Speech / Duration of All Speech

False Alarm = Missed Non-Speech / Duration of All Speech

Other metrics
(CHIL)

SDER = Missed Speech / Duration of All Speech

NDER = Missed Non-Speech / Duration of All Non-Speech

Performance, cont.

UPC primary systems

	NIST / Missed Spkr / False Alarm *SDER / NDER*	
	confmtg	lectmtg
mdm	**5.70** / 3.5 / 2.2 ? *3.5 / 39.5*	**8.62** / 2.8 / 5.8 *2.8 / 44.4*
sdm	**5.51** / 3.1 / 2.4 *3.1 / 42.1*	**9.40** / 1.6 / 7.8 *1.6 / 58.9*

Performance, cont.

Other UPC systems

	NIST / Missed Spkr / False Alarm *SDER / NDER*	
	contrastive confmtg 10 features	post-eval confmtg SVM
mdm	**5.70** / 3.6 / 2.1 *3.5 / 37.1*	---
sdm	**5.51** / 3.3 / 2.3 *3.3 / 40.1*	**4.72** / 0.8 / 3.9 *0.8 / 68.8*